Vol. 14 No. 1 天 文 研 究 与 技 术 第 14 卷 第 1 期
Jan., 2017 ASTRONOMICAL RESEARCH AND TECHNOLOGY 2017 年 1 月

CN 53-1189/P ISSN 1672-7673

新疆天文台数据中心建设与应用*

张海龙[1,3]，王 杰[1]，王万琼[1]，聂 俊[1]，冶鑫晨[1]，朱 艳[1,2]，托乎提努尔[1]

(1. 中国科学院新疆天文台，新疆 乌鲁木齐 830011；2. 中国科学院大学，北京 100049；
3. 中国科学院射电天文重点实验室，江苏 南京 210008)

摘要：经过多年观测，新疆天文台积累了大量珍贵的观测数据。如何高效管理快速增长的海量观测数据，如何实现全波段观测数据的融合及未来更大口径望远镜产生的海量数据归档与发布是目前新疆天文台急需解决的问题。新疆天文台数据中心面向天文学科研需求提供天文科学数据服务和基础设施，由新疆天文台计算机技术室负责运行与维护。数据中心以最新的虚拟天文台相关标准为基础建设，基本服务包括：为天文观测项目提供数据保存、管理和发布服务；为有价值的天文科学数据及二次处理后的数据提供长期保存和访问服务；为用户提供科学数据归档与发布服务及相关技术支持。数据资源主要有：新疆天文台 25 m 射电望远镜的脉冲星观测数据；活动星系核观测数据以及分子谱线观测数据。提供的数据服务有：PPMXL 星表锥形检索；海量星表数据在线交叉认证；统一内容描述符信息查询等服务。主要介绍新疆天文台数据中心建设及其提供的服务，如何使用已发布的服务在线检索数据及利用标准虚拟天文台工具实现数据操作。

关键词：数据中心；虚拟天文台；数据发布；检索

中图分类号：TP3-05 **文献标识码**：A **文章编号**：1672-7673(2017)01-0094-09

当今世界各国都加大了对先进观测设备的研究和投入，仅 2008 年，我国就有 3 座大型天文望远镜建成或奠基，分别是每夜光谱观测量上万的国际上最具威力的大天区面积多目标光纤光谱天文望远镜(Large Sky Area Multi-Object Fiber Spectroscopy Telescope, LAMOST)[1]；即将投入使用的 500 m 口径球面射电望远镜(Five-hundred-meter Aperture Spherical Radio Telescope, FAST)[2]；可在厘米、分米波段上同时实现以高空间、高时间和高频率分辨率观测太阳动力学过程的射电频谱日像仪。国际上，平方千米阵列(Square Kilometer Array, SKA)的几个先驱阵列，ASKAP[3]、LOFAR[4]、MeeRKAT[5]等已经产生海量的天文数据。

计算机网络技术的快速发展为海量天文数据的共享及虚拟天文台技术提供了有力的支持[6]。海量天文数据的访问要求传统的数据访问应用或服务能够应对前所未有的海量天文数据的压力[7]。在文档库中检索关键词的全文检索技术目前已经取得了较大进展[8]。数据的指数级增长给数据库查询提出了更严峻的挑战，为了在海量天文数据中快速定位目标文件，需要研究海量天文数据检索、挖掘、交叉认证等技术[9]。在天文界，FITS[10-11]标准文件包括两类数据格式，一类是 ASCII 码，另一类是二进制格式，这两种数据格式都受两条基本限制所制约。FITS 文件不适合并行处理及对于不同类别数据合并计算后的元数据不够精确，这两条限制决定了天文数据检索的特殊性。

以脉冲星观测为例，澳大利亚帕克斯(Parkes)64 m 射电望远镜是南半球唯一的大型射电望远镜，具有得天独厚的地理优势，能够长时间观测银河系中心和南银纬的天体，现在能观测到一千多颗脉冲星[12]。目前已知的所有(射电)脉冲星，超过一半是帕克斯 64 m 望远镜发现的。帕克斯脉冲星数据库

* 基金项目：国家自然科学基金 (U1531125)；国家重点基础研究发展计划 (973 计划) (2015CB857100)；中国科学院青年创新促进会；中国科学院西部之光项目 (XBBS201325)；中国科学院天文台站设备更新及重大仪器设备运行专项经费；天文学科技领域云项目 (XXH12503-05-05) 资助.

收稿日期：2016-04-01；修订日期：2016-04-19

作者简介：张海龙，男，博士. 研究方向：数据密集型研究. Email：zhanghailong@xao.ac.cn

拥有近 20 万个数据文件[13]，是 1991 年以来观测的脉冲星搜寻数据和脉冲星到达时间数据，这些数据可以公开免费获取。对于到达时间模式的观测，除了原始数据外，脉冲星数据库还包含相应的"完全压缩"后的数据，即在时间、偏振和频率上完全积分后的数据，这给使用者提供了另一种选择：下载压缩后的文件比下载原始数据花更少的时间。澳大利亚帕克斯 64 m 射电望远镜观测的数据格式为 PSRFITS 格式、Timer 格式。帕克斯的数据发布网站上提供了基于 RA、DEC、Cone Search 等服务①。

新疆南山 25 m 射电望远镜脉冲星多通道消色散观测系统从 2000 年正式开始运行，截止到目前，常规地对 300 颗脉冲星进行了 9~11 年的到达时间观测，研究了脉冲星周期跃变、星际闪烁、时间噪声、脉冲星自行、轮廓模式变化等。2006 年开发了单脉冲观测模式，研究了巨脉冲、强脉冲、暂现源和磁星的射电辐射。这些研究项目的观测由多人完成，十几年的观测数据存储分散，原始数据往往保存在个人及研究组的电脑中，目前部分数据已过了保护期。建设新疆天文台的数据发布与检索系统，使脉冲星数据可以对全世界的科研人员开放共享。

很多科学研究成果来源于对历史数据的重新处理。帕克斯多波束巡天的数据后来被重复处理了多次，又发现了大约 30 颗脉冲星和 10 颗暂现源。再如，对大麦哲伦云的脉冲星巡天数据的重复分析处理，发现了一个很强的持续时间很短的射电暴，它可能来自银河系外。很多研究成果也来源于几年甚至几十年长时间观测的数据积累，比如脉冲星的时间噪声、脉冲星的周期跃变等。但是这些数据的再处理及新目标的发现均需要完整的数据归档及数据发布系统支撑，所以建设系统、完整的新疆天文台数据归档与发布平台可以在更好地适应国际天文学发展的趋势、提高数据的利用率、产生更大的科研价值等方面打下坚实的基础。

1　数据中心建设

新疆天文台数据中心访问地址：http://data.xao.ac.cn(或 http://vo.xao.ac.cn)，两个域名分别指向两台互为备份的数据存储服务器，一台位于新疆天文台本部，一台位于南山观测站。数据中心主页如图 1。

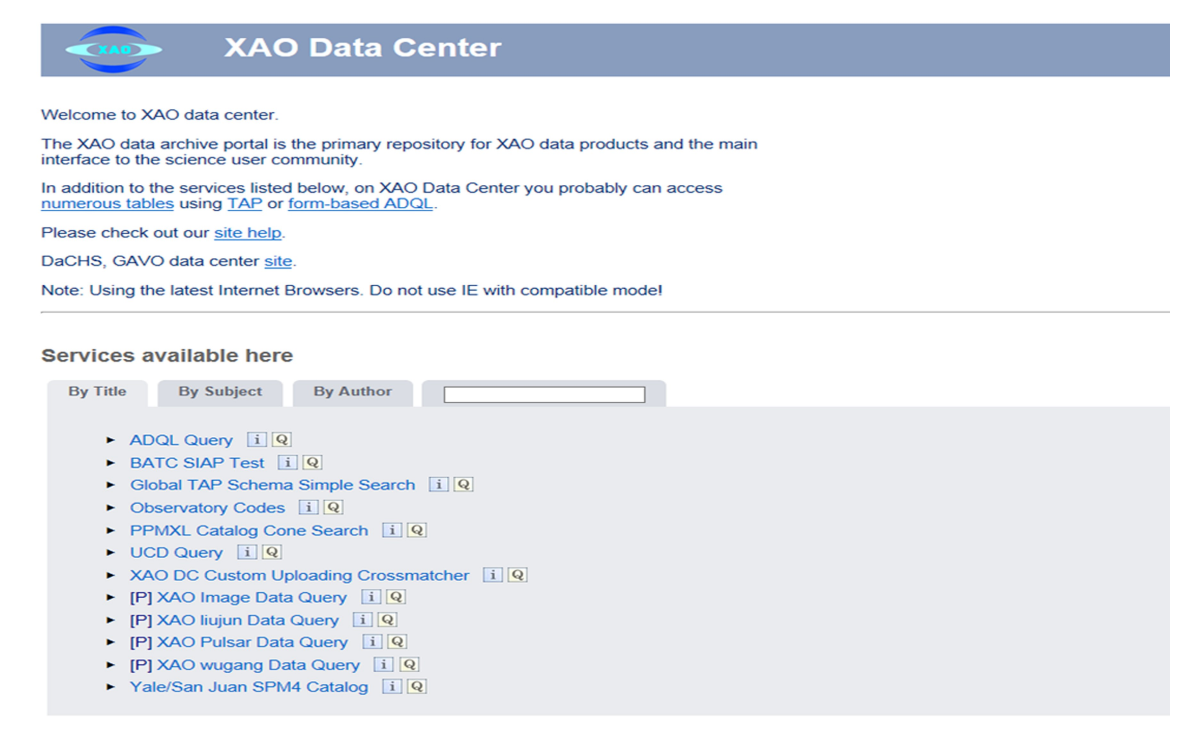


图 1　新疆天文台数据中心主页

Fig. 1　The homepage of XAO data center

① https://datanet.csiro.au/dap/public/atnf/pulsarSearch.zul

两台数据服务器的配置信息如表 1，分别提供 12 核 24 线程 2.2 G 中央处理器两颗、48 TB 存储空间、64 GB 内存。目前可满足新疆天文台数据归档与发布的需要。

表 1 数据服务器配置信息

Table 1 Configuration information

配件	规格参数	数量
中央处理器	Intel （R） Xeon （R） CPU E5-2692 v2 @ 2.20 GHz	2
内存	8 GB	8
硬盘(OS)	300 GB	2
硬盘(DATA)	4 TB	12
主板	Intel Corporation C600/X79 series chipset	1
附加网络接口	IB Card 56 Gbps	1
机箱	2 U	1

2 数据中心服务

2.1 天文数据查询语言

天文数据查询语言②(Astronomical Data Query Lauguage，ADQL)是一种类似 SQL 的语言，用来检索天文星表或者其他以表形式存在的数据集。

图 2 为天文数据查询主页面，服务对应网址 http://data.xao.ac.cn/_system_/adql/query/form，通过主页面可实现对新疆天文台已发布的所有数据进行查询操作。

实例 1：select name,code from obscode.data where name like ‘%Nanshan%’ and code like ‘%1’，可实现在已发布的观测台站坐标查询服务中查找名字包含“Nanshan”字符串的台站。

实例 2：select * from ppmxl.main where 1=CONTAINS（POINT(‘ICRS’，raj2000，dej2000)，CIRCLE(‘ICRS’，10.684 792 9，41.269 065 0，0.1))，可实现检索 PPMXL 星表中以 M31(10.684 792 9，41.269 065 0)为中心，0.1°半径天区内所有的目标。

2.2 锥形检索

锥形检索③(Cone Search)是在指定半径的一个锥形天区中的一种检索。作为虚拟天文台体系中第 1 个数据访问协议，锥形检索力求简便，从而使数据发布者能快速将其实现成为一个有力的搜索工具。一个锥形检索服务必须支持带有 3 个输入参数的 HTTP GET 请求，并做出反应。请求中 3 个必要参数分别指赤经(RA，J2000)、赤纬(DEC，J2000)以及检索半径(SR)。3 个参数都以十进制为单位，通过一个可选的冗余参数，用户可以指定服务返回表格形式的列数。

新疆天文台发布的 PPMXL 星表数据锥形检索服务网址 http://data.xao.ac.cn/ppmxl/q/cone/form。如图 3，在 Position/Name 栏输入 M31，在 Search radius 栏输入 6，可查询以 M31 为中心，6″为半径在 PPMXL 星表中有多少个目标，输入完成后点击 GO 按钮可以看到相应的检索结果。

2.3 表访问协议

表访问协议④(Table Access Protocol，TAP)定义了一个服务协议访问表数据，包括天文目录以及通用数据库表。提供访问数据库和表的元数据以及实际的表数据。TAP V1.0 协议支持多种查询语言，包括天文数据查询语言和正在开发中的参数化查询语言(Parameterised Query Language，PQL)，它还支持同步和异步的数据查询操作，对天文数据查询语言中的空间扩展及索引查询提供特殊支持。多目标

② http://www.ivoa.net/documents/latest/ADQL.html
③ http://www.ivoa.net/documents/latest/ConeSearch.html
④ http://www.ivoa.net/documents/TAP/

查询功能允许查询任意大型天文目标列表，并提供了简单的空间交叉认证的功能。更复杂的分布式交叉认证功能可以通过分布式查询实现。

ADQL Query

data.xao.ac.cn/__system__/adql/query/form

Home
Help
Service info
Related
Tables available for ADQL
Metadata
Identifier
ivo://xaovo/__system__/
Description
An endpoint for submitti
Keywords
Virtual observatory
Creator
Created
2008-09-20T12:00:00Z
Data updated
2015-04-01
Reference URL
Service info
Try ADQL to query our data.
XAO web site.

ADQL Query

On this page, you can use ADQL to query some of our tables. This is mainly for dabbling; use TAP for larger jobs (e.g., using TAPHandle within your browser).

To learn what ADQL is or for further information on this implementation, see the service info.

ADQL query: select * from ppmxl.main where 1=CONTAINS (POINT('ICRS',raj2000,dej2000), CIRCLE ('ICRS', 10.6847929 , 41.2690650,0.1))

A query in the Astronomical Data Query Language

Timeout after [s]: 5

Seconds until the query is aborted. If you find yourself having to raise this beyond 200 or so, please contact the site operators for hints on how to optimize your query

Output format: HTML

Go　　[Result link]

To protect your nerves, the server inserts a TOP 2000 phrase unless you give a limit yourself. Thus, if you hit the 2000 record limit and want to override it, you can specify your limit yourself (like SELECT TOP 20000...).

There is a fixed limit to 100000 rows on this service. If this bugs you, use TAP.

正在连接...

图 2　天文数据查询界面

Fig. 2　The Interface of ADQL

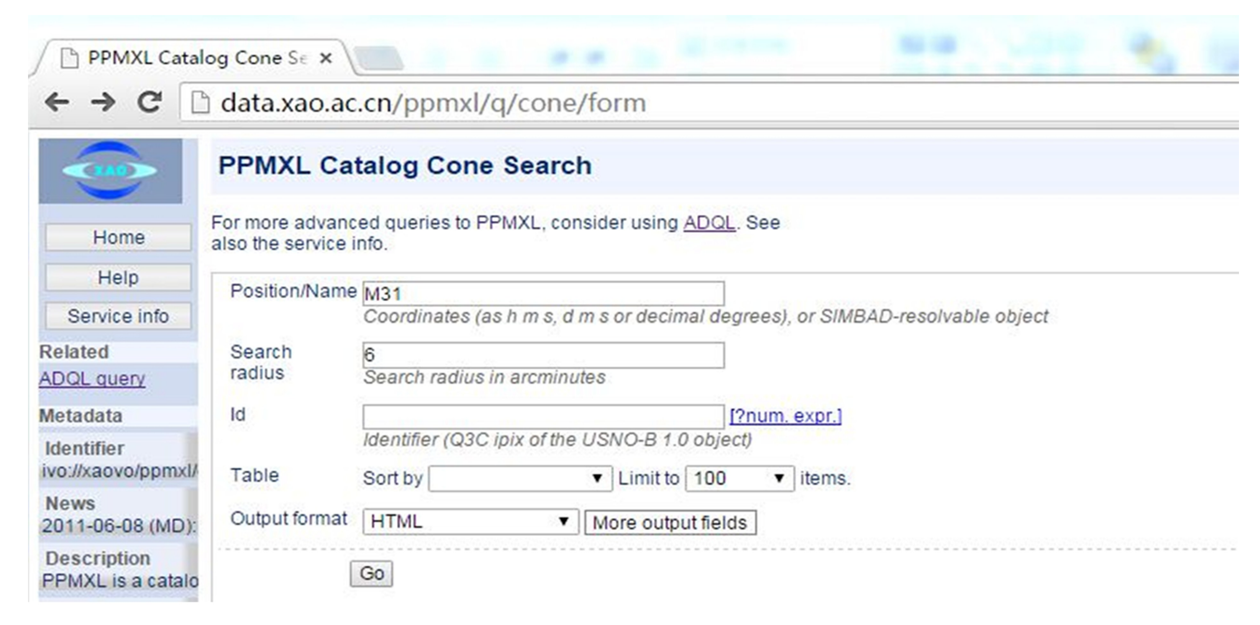


图 3　PPMXL 星表数据查询

Fig. 3　PPMXL catalog query

新疆天文台发布的表访问协议查询服务如图 4，网址 http://data.xao.ac.cn/glots/q/plain/form。服务可以实现已注册到国际虚拟天文台联盟(International Virtual Observatory Alliance，IVOA)中的表及字段数据的直接查询。

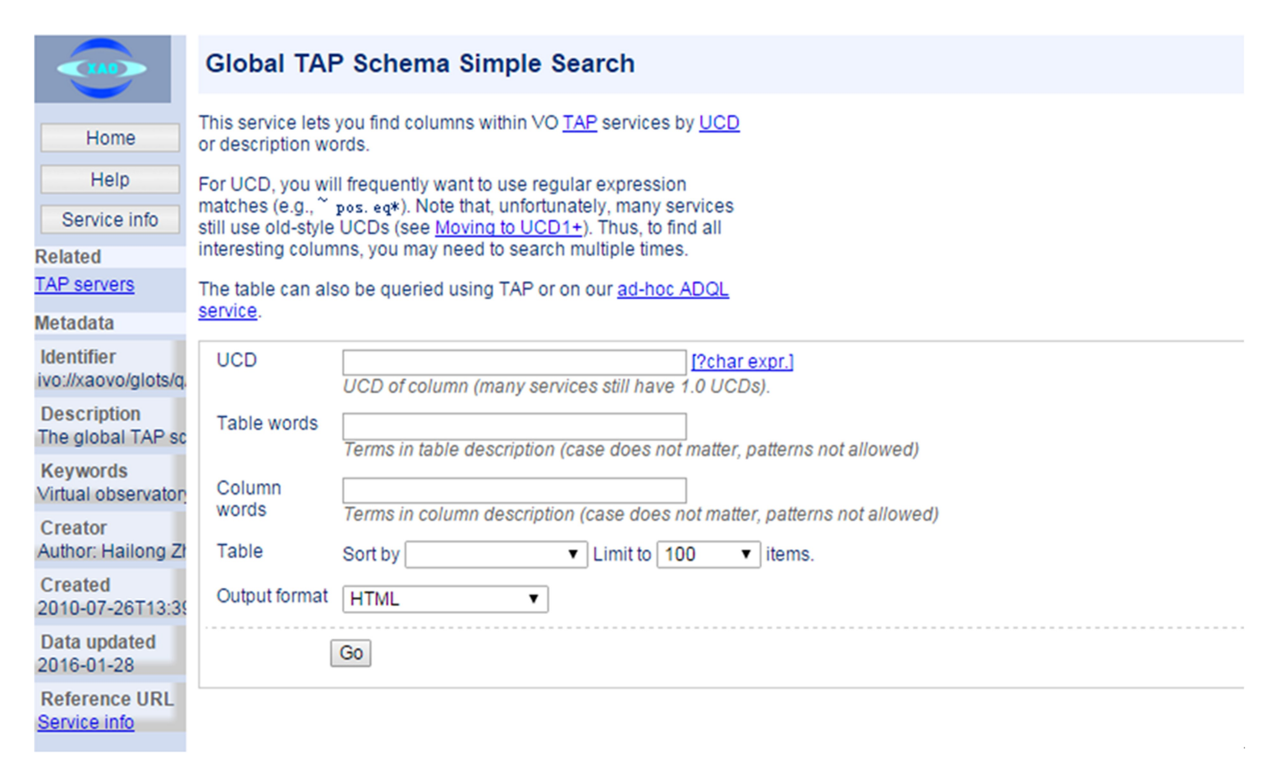


图 4 表访问协议查询服务

Fig. 4 TAP Service

2.4 简单应用程序消息传递协议

简单应用程序消息传递协议⑤(Simple Application Messaging Protocol，SAMP)可以实现天文软件间的互操作和数据交换。建立一个独立的工具，试图满足所有用户的需求是不切实际的，任何工具软件都不是万能的。但是可以通过协议实现，协议可以约束软件的行为与接口，实现消息的正常传递。简单应用程序消息传递协议是一个高效利用有限的资源，使个人工具更好地协同工作的协议。此协议的特性是定义常见的文件格式完成不同应用程序之间的数据交换，重要组件是一个消息传递系统，它允许应用程序共享数据和利用彼此的功能，支持桌面和网络浏览器间的应用程序通信，形成一个更一般的消息传递框架以满足后续消息传递的需求。

新疆天文台发布的数据服务均支持简单应用程序消息传递协议，可将查询结果直接传递给标准的虚拟天文台工具，进行数据挖掘、可视化等相关操作。

2.5 新疆天文台数据中心在线交叉认证服务

2.5.1 交叉认证描述

天文交叉认证是实现多波段数据融合的基础[14]，是加深对天体的认识、促进天文学新发展的关键一步，同时也是大型望远镜在观测前选源阶段进行星体类型识别的核心技术。交叉认证是指两个不同波段的星表数据，以第 1 个星表源的位置为中心，对该星表中的每一个源都在另一个星表中寻找对应体，若两个源之间的球面角距离小于某一特定值，就认为这两个源是同一天体。

2.5.2 交叉认证原理

以图 5 中的两点 A、B 为例，它们分别来源于星表 A 和星表 B，设其坐标分别为(α_1, δ_1)、$(\alpha_2,$

⑤ http://www.ivoa.net/documents/SAMP/

δ_2)，它们之间的球面角距离可以按如下步骤计算：

(1) 若 $|\alpha_1-\alpha_2|\leqslant 180°$，　$\angle ANB=|\alpha_1-\alpha_2|$，

(2) 若 $|\alpha_1-\alpha_2|>180°$，　$\angle ANB=360°-|\alpha_1-\alpha_2|$.

根据球面余弦定理：

$$\angle AOB=\cos\angle AON\cos\angle BON+\sin\angle AON\sin\angle BON\cos\angle ANB$$

$$=\sin\delta_1\sin\delta_2+\cos\delta_1\cos\delta_2\cos(\alpha_1-\alpha_2),$$

$$d=\angle AOB=\arccos[\sin\delta_1\sin\delta_2+\cos\delta_1\cos\delta_2\cos(\alpha_1-\alpha_2)].$$

当 A、B 两点之间角距离很小时，$\delta\approx(\delta_1+\delta_2)/2$，

所以有 $d^2=[(\alpha_1-\alpha_2)\cos\delta]^2+(\delta_1-\delta_2)^2$。

认证成功的判断公式：

$$d=\sqrt{[(\alpha_1-\alpha_2)\cos\delta]^2+(\delta_1-\delta_2)^2}\leqslant 3\sqrt{r_1^2+r_2^2} \quad (1)$$

其中，r_1 和 r_2 为两个星表的误差半径，也就是当两点之间的距离满足上式时，可以认为两点认证成功，互为匹配的对应体。

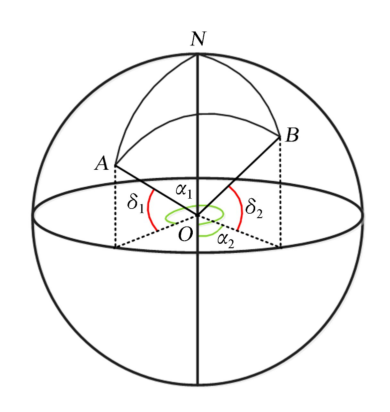


图 5　球面两点交叉认证原理

Fig. 5　The principle of cross validation for two spherical points

2.5.3　交叉认证实现

新疆天文台发布的在线交叉认证服务如图 6，目前可以实现本地文件及远程统一资源定位符(Uniform Resource Locator，URL)方式上传带有 RA(ucd：pos.eq.ra)、DEC(ucd：pos.eq.dec)信息的星表数据文件。选择目标星表，可以是新疆天文台发布的所有数据中任何一张表，最后选择所要匹配的搜索半径，由于单台观测设备存在误差，所以根据自己星表数据产生设备的具体指标输入搜索半径。之后点击 GO 按钮，可以得到认证结果。

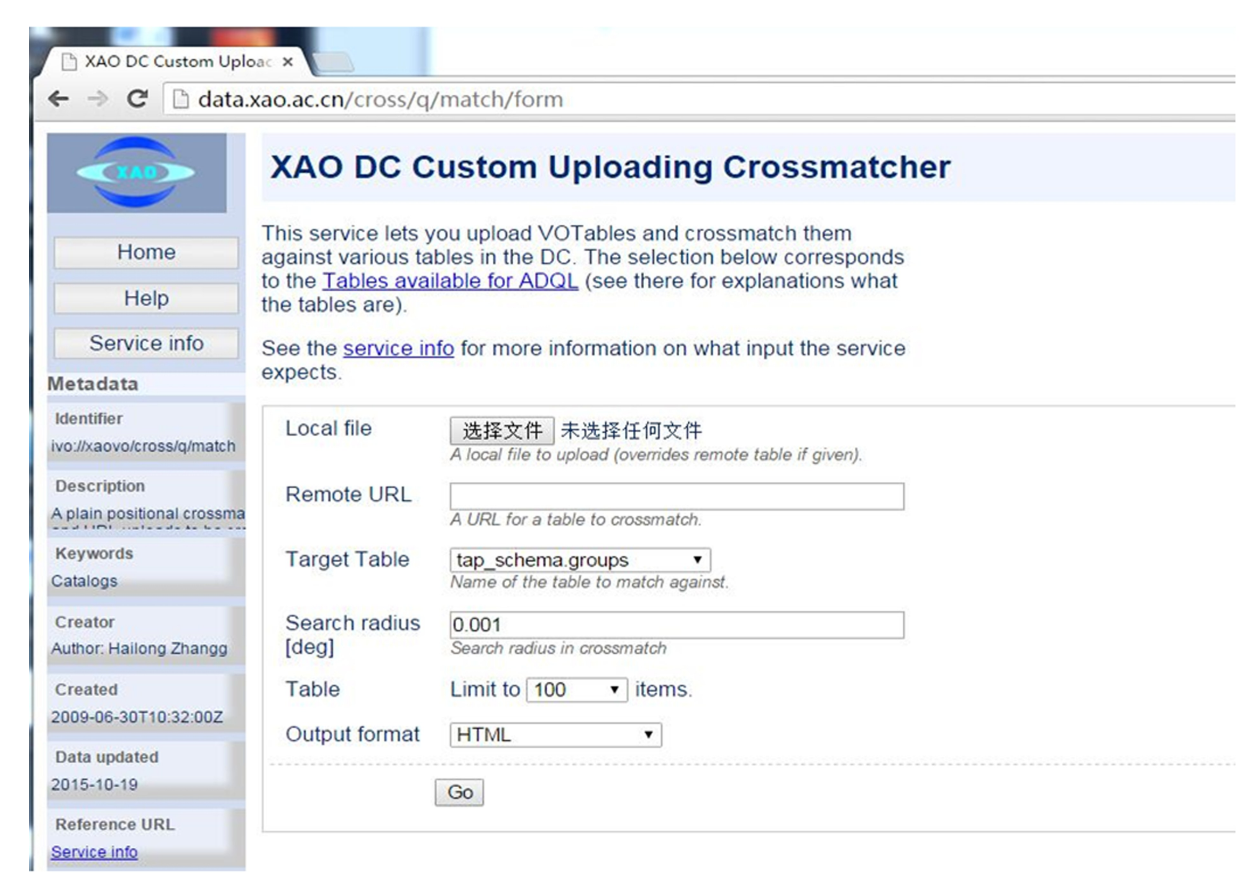


图 6　交叉认证服务

Fig. 6　Cross validation service

2.5.4 参考实验

利用新疆天文台发布的 PPMXL 星表为测试目标，以 URL：http://data.xao.ac.cn/static/cross-match (4000 星表记录)为输入条件，选择 PPMXL.MAIN(近 10 亿条记录)星表，0.001°为半径，在毫秒量级时间内可以得到匹配结果，大概 757 条。

2.6 新疆天文台观测数据服务

数据中心已基本完成对南山 25 m 射电望远镜观测产生的脉冲星、短时标光变(Intra-Day Variability, IDV)、分子谱线等科研数据的在线存储、归档、备份，并向全台相关研究人员提供数据的发布与存储等服务。以脉冲星数据为例，脉冲星数据检索网址 http://data.xao.ac.cn/pul/pulsar/q/form，脉冲星数据实现了锥形检索与基于元数据信息的多目标约束检索。

以脉冲星数据检索系统为例，打开检索页面并在“Target Object”中选“J0332+5434”，在“Data_obs”中填入“~ *2014 *”，在“Output format”中选择“HTML”。以 HTML 方式输出时可以预览脉冲轮廓及其它相关信息。如果期望得到更多返回结果，可以调整 LIMIT TO 值。在“Preview”中点击“[Preview(pav-DFTp)]”可以查看相应脉冲轮廓大图。同时在输出左上角有一个“Send via SAMP”按钮，可将检索结果送给支持简单应用程序消息传递协议的虚拟天文台工具，利用 TOPCAT 等软件可接收数据检索结果并进行可视化操作。具体查询及使用方法详见文[15]。

3 结果输出格式

数据中心支持多种数据格式输出，可选择 HTML、FITS Table、CSV、JSON、VOTable、tar 等格式的数据输出，如果数据量较大且文件个数较多时可选择 tar 格式，直接打包下载所有满足条件的数据。

因为浏览器响应需要时间，数据量大时会导致无响应。Limit to 字段约束结果输出的行数，可根据需要调整相关数值。More output fields 中用户可选择更多的输出字段，从而得到更丰富的数据信息。

4 虚拟天文台工具查询

数据中心的脉冲星数据、PPMXL 星表锥形检索等服务已完成 IVOA 注册，通过标准虚拟天文台工具，可以检索到我台已发布的数据信息，利用 TOPCAT 锥形检索工具可查找已注册到 IVOA 的服务中带有“xao”字符的服务，(新疆天文台缩写为 XAO，所以我台发布的服务均可通过“xao”字符查找)，具体查询及使用方便详见文[15]。

5 总 结

新疆天文台数据中心已基本具备数据归档和发布功能，并以虚拟天文台协议为基础实现了新疆天文台历史数据的归档与发布。目前已归档脉冲星数据 32 390 条记录，最大文件 1 GB，针对脉冲星数据建立了以锥形检索、观测目标等为基础的多约束目标检索；针对短时标光变、恒星形成实现了锥形检索服务。以并行计算技术为基础发布的海量数据在线交叉认证服务可实现高速在线交叉认证，并且提供了本地数据上传及远程统一资源定位器两种方式，上传后数据可与数据中心已发布的星表数据及其它数据进行交叉认证，测试结果显示以 4 000 目标匹配 10 亿目标，计算返回时间为毫秒量级。数据中心的天文数据查询服务使用户可以通过天文数据查询语言查询到数据中心发布的所有允许查询的数据；UCD 查询服务可以实现统一资源定位符内容的查找；天文台代码查询服务可以实现全世界各天文台站的坐标相关信息查询。

新疆天文台数据中心的数据挖掘与可视化相关操作还有待进一步完善。

致谢：感谢天文学科技领域云项目成员对新疆天文台数据中心建设的支持。数据中心测试与数据的预处理在新疆天文台 Taurus 高性能计算系统上完成。

参考文献：

[1] 崔向群. LAMOST 项目及进展 [J]. 天文学进展, 2001, 19(2): 123-128.
Cui Xiangqun. LAMOST project and its progress [J]. Progress in Astronomy, 2001, 19(2): 123-128.

[2] 南仁东, 李会贤. FAST 的进展——科学、技术与设备 [J]. 中国科学, 2014, 44(10): 1063-1074.
Nan Rendong, Li Huixian. Progress of FAST in science, technique and instrument [J]. Science China, 2014, 44(10): 1063-1074.

[3] Johnston S, Taylor R, Bailes M, et al. Science with ASKAP [J]. Experimental Astronomy, 2008, 22(3): 151-273.

[4] Mol J D, Romein J W. The LOFAR beam former: implementation and performance analysis [J]. Lecture Notes in Computer Science, 2011, 6853: 328-339.

[5] Jonas J L. MeerKAT-the South African array with composite dishes and wide-band single pixel feeds [J]. Proceedings of the IEEE, 2009, 97(8): 1522-1530.

[6] 崔辰州, 赵永恒. 中国虚拟天文台体系结构 [J]. 天文研究与技术——国家天文台台刊, 2004, 1(2): 140-151.
Cui Chenzhou, Zhao Yongheng. Architecture of Chinese virtual observatory [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2004, 1(2): 140-151.

[7] 宋雯斐. 全文检索研究与系统实现 [D]. 北京: 北京师范大学, 2004.

[8] 喻宏勇, 赵铁军, 郑德权, 等. 一种面向文件的高效检索方法 [C]// 中国中文信息学会二十五周年学术会议. 2006.

[9] 赵青, 孙济洲, 于策, 等. 面向海量数据的并行天文交叉证认 [J]. 计算机应用, 2010, 30(8): 2096-2060.
Zhao Qing, Sun Jizhou, Yu Ce, et al. Parallel massive data oriented astronomical cross-match [J]. Journal of Computer Applications, 2010, 30(8): 2096-2060.

[10] 崔辰州, 李文, 于策, 等. FITS 数据文件的检索和访问 [J]. 天文研究与技术——国家天文台台刊, 2008, 5(2): 116-123.
Cui Chenzhou, Li Wen, Yu Ce, et al. Search and location of FITS data files [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2008, 5(2): 116-123.

[11] Wells D C, Greisen E W, Harten R H. FITS: a flexible image transport system [J]. Astronomy & Astrophysics Supplement Series, 1981, 44: 363-370.

[12] Manchester R N, Lyne A G, D'Amico N, et al. The parkes Southern pulsar survey-I. observing and data analysis systems and initial results [J]. Monthly Notices of the Royal Astronomical Society, 1996, 279(4): 1235-1250.

[13] Hobbs G, Millera D, Manchester R N. The Parkes observatory pulsar data archive [J]. Publications of the Astronomical Society of Australia, 2011, 28(3): 202-214.

[14] 赵青. 面向海量数据的高效天文交叉证认的研究 [D]. 天津: 天津大学, 2010.

[15] Zhang Hailong, Markus Demleitner, Wang Na, et al. Data retrieval from Xinjiang Astronomical Observatory's pulsar data archive [J]. Astronomical Research & Technology, 2016, 13(4): 473-480.

Construction and Application of the Data Center in Xinjiang Astronomical Observatory

Zhang Hailong[1,3], Wang Jie[1], Wang Wanqiong[1], Nie Jun[1], Ye Xinchen[1], Zhu Yan[1,2], Tohtonur[1]

(1. Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi 830011, China, Email: zhanghailong@xao.ac.cn;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing 210008, China)

Abstract: Xinjiang Astronomical Observatory (XAO) has accumulated massive valuable observational data after years of observation. Problems including how to efficiently manage the rapid growth of the massive observational data, how to achieve the integration of the whole band observations as well as archive and publish massive data produced by larger diameter telescopes in future are extremely urgent and hence need to be solved in a timely manner. XAO data center is a scientific data service infrastructure designed to meet the demands of scientific research in astronomy. ICT is in charge of the implementation and maintenance of this project. This data center is constructed based on the relevant standards of virtual observatory. Its basic services include: providing services of data storage, management and release for astronomical observation programs, offering services of long-term preservation and access for valuable astronomical data and processed secondary data, offering scientific data archiving, release and technical support for users. Its main data resources include pulsar observation data, observed data of active galactic nuclei and molecular spectrum, all of which are obtained by the 25m Radio Telescope in XAO. Data services provided by this center include PPMXL catalog data online massive cone search, cross validation, UCD information query services and so on. This paper will mainly focus on introducing the construction of the XAO data center, its services, and how to use these services to retrieve data or use standard VO tool to realize data operation.

Key words: Data center; Virtual Observatory; Data releases; Index